学校编码：10384　　分类号密级

学号：X2011230712 UDC

工　程　硕　士　学　位　论　文

互联网信息采集系统的设计与实现

Design andImplementation of Internet Information AcquisitionSystem

白涛

指 导 教 师：　王 备 战 教 授

专 业 名 称：　软　件　工　程

论文提交日期：　2014 年 6 月

论文答辩日期：　2014 年 7 月

学位授予日期：　　年　　月

指导教师：

答辩委员会主席：

2014 年 7 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（）课题（组）的研究成果，获得（）课题（组）经费或实验室的资助，在（）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年月日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（）1.经厦门大学保密委员会审查核定的保密学位论文，于

年　月　日解密，解密后适用上述授权。

（）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年月日

摘要

互联网信息采集是政企部门信息采集管理的重要工作，对于信息收集并进行资源管理建设方面作用明显。随着我国互联网信息服务管理改革的不断深入，互联网资源的收集、整合与管理引起了广泛关注。过去，政企部门人工采集的方式消耗了大量的人力、物力以及财力，考核结果也不便于保存，难以适应于信息化时代的需要。当今社会，随着现代社会信息网络的不断扩充，互联网技术已经遍布到生活的各个角落，许多基于互联网的应用程序也日益增长起来，极大地方便了我们的工作和生活以及学习。所以，开发一套科学的互联网信息采集系统，帮助政企部门进行绩效管理，成为了当前急待解决的问题。

论文对目前互联网信息的采集效益进行了调查研究，分析了智能化采集互联网资源的重要意义，对目前网络系统常用的架构模式进行了介绍，并结合本单位实际需要设计开发了互联网信息采集系统。本系统经过 twitter 授权，通过调用 twitter 外部接口来获取信息资源。从需求分析入手，采取了模块化设计思想，对系统的角色和功能进行了划分。

本系统实现了实时消息抓取、用户习惯分析、twitter 用户检索、twitter 内容检索、互联网全网检索等功能，秉着人性化的设计理念，设计了友好的用户界面，并在安全性上加大了考虑，针对目前常见的网络攻击设计了防范措施，保证了系统的稳定运行。系统实现有效的满足了政企部门进行信息采集考核工作的需求，并对如何进行科学的、快速的、有效的采集资源起到了一定的促进和推动作用，对于同类系统的设计和开发也具有一定的借鉴作用。本系统在实施应用中取得了一定的效果，达到了预期目标。

关键词：互联网；信息采集；Twitter

Abstract

Internet information collection is an important work of government and enterprise sector information collection management, collection of information and resources for the management of construction effect is obvious. With the deepening of the reform of China's Internet Information Services Manager, collection, integration and management of Internet resources wide attention. In the past, government and enterprise sectors artificial way of collecting consumes a lot of manpower, material and financial resources, assessment results and certainly not easy to save, is difficult to adapt to the needs of the information age. Today's society, with the amplification of modern social information network, the Internet technology has spread to every corner of life, many Internet-based applications are also growing up, which greatly facilitated our work and life, and learning. Therefore, the development of a scientific Internet information collection system, to help government and enterprise performance management department, became the current problems need to be solved.

This dissertation is currently gathering information on the benefits of the Internet were investigated, analyzed the importance of intelligence collection of Internet resources on the current network architecture commonly used modes are introduced. The system has been authorized twitter, twitter by calling an external interface to access information resources. Starting from requirements analysis, to take a modular design concept of the role and functions of the system has been divided, using PHP as a development language, combined with the current more popular and a higher level of security MYSQL database design to achieve a set of browser-based Internet information collection system.

The system enables real-time news crawl, user habits analysis, twitter users to retrieve, twitter content retrieval, the whole network of Internet search functions, is holding humanized design concept, design user-friendly interface, and increase in security the consideration for the current common network attacks preventive measures designed to ensure the stable operation of the system. Effective

implementation of the system to meet the government-enterprise sector information collection needs assessment work, and how to conduct scientific, rapid and effective collection of resources has played a certain role in promoting and facilitating, for the design and development of similar systems also have certain reference. This system has achieved some results in the implementation of applications, to achieve the desired goals.

Keywords: Internet; Information Collection; Twitter

目录

Contents

第一章绪论

1.1 研究背景和意义

随着我国科学技术的不断发展与革新，新的科技革命日趋白热化。大大地影响了人类社会的各个领域，在日常生活和社会生产中，互联网技术的革新占据了首要地位，它在人们没有察觉的情况下悄悄地潜入了人类社会，而且越来越依赖它。其中，利用计算机迅速可靠、节省资源的特点对互联网资源的绩效管理是各个政企部门想要发展顺畅的优先选择。而信息的择优采集则是绩效管理中的重中之重，且这部分的内容是我们系统的研究方向和基石，当然必不可少，对于管理者和采集者，实时抓取与收集互联网信息尤为重要。但是一直以来，政企部门人员使用人工手动采集的方法来收集信息，更有甚者采用纸质的方法管理信息，这样首先在安全性方面来说，就不过关，保密性非常差，其次对于自然环境也是极不友好的，我们应该在保护森林覆盖率的长远计划中尽自己的一份力，最后这在信息的及时更新上也是非常不可取的，同时大大降低了工作人员的工作效率，并且增加了工作量。

作为互联网应用的一部分，通过互联网采集系统对信息进行科学的采集管理，使得获取信息资源更加的迅速、方便，而且与旧时的方法相比，更加的环保、准确以及安全，以后更新信息或者修改内容都比较便捷。这样不仅节省了资料的浪费而且提高了工作人员的效率，使得各大单位的企业文化得到提升，人们的认知水平也更上一个台阶，我国企业想在国际上获得更高的评价也是指日可待。

近几年来，为了响应我国“科技强国”的战略手段，非常多的单位、高校以及政府单位都对从互联网上获取第一手资料投入了相当一部分的人力、财力，而且不同程度上的促进了生产，各界也是连连称赞[1]。其成效包括：各个学校纷纷采用了 Internet 技术，来开阔学生的眼界，看看不同于我们的世界；很多企业单位也都紧跟着互联网技术迅速发展；各个政府单位都设立了自己的门户网站，方便倾听民众的声音，管理者也都颇为重视。

但是随之而来的就是信息量的不断暴增。各种类型的网站不断地架设，使得采集自己想要的，最新的信息变得异常的艰难。此时，垃圾软件以及垃圾信息污

染了整个互联网，工作人员的工作效率非但没有提示反而更加的杂乱，项目进度也有可能会受到牵连，稍有闪失，就会给单位带来极大的损失而不是收益。所以好的系统软件的需求迫在眉睫。

互联网信息采集系统是在这样的环境下开发出来的。该系统有效地改善了政企部门采集信息中会遇到的以上问题。它以一种更为科学的管理方法使得互联网信息的收集逐渐变得自动化、一体化，对于前面提到在信息收集中存在的信息泛滥、难辨真假、更新缓慢等问题都得到很好的解答。为明确信息采集的研究方向提供了很大的帮助。

1.2 国内外研究现状

其实，不只我国在信息采集管理软件方面进行研究，早在我们之前国外的许多科研机构就在这方面进行了深入的探索，而且探索结果也是不错的[2]。但是我们是具有中国特色的社会主义国家，当然我们的研究方向也需具备中国特色，而且国外的多数是英文版的，这对于信息采集的工作人员来说，不免又多要求了一份。对于后续的开发应用增加了难度。

在国内，随着互联网的迅猛发展，从前独立单一的手工采集信息的方法早已不能满足我们的需求。进入 21 世纪后，各大政企单位以及高校如雨后春笋般的相继创立了自己的门户网站，而且浏览器/服务器等结构模型也越来越趋于平稳发展，操作系统也日益流畅，这让软件程序变得更容易上手，不仅满足了对数据信息实时采集、抽取、挖掘、处理的需求，而且让信息及时地更新变得更加容易。

虽然现在很多企业都从市面上采购了一些采集系统，但是这些系统都是针对各个单位各自的特点设计的，除此之外，这些系统使用方式较为分散，用户界面比较复杂。针对以上问题，论文设计研发的互联网信息采集系统提出了相应的解决方案，实现了采集信息的智能化、现代化，并通过测试，已经用于一些单位的日常工作使用，并取得良好的效果。

1.3 研究内容

论文以现下互联网的基本存在的问题为主，针对各个政企部门及高校的共同需求，研发了这套互联网信息采集系统，该系统具备各类数据的自动化处理的功

能。互联网信息采集系统主要是通过 twitter[3]授权认证，实时抓取时下最新鲜的推文，并对用户的习惯进行了系统的分析，还可以进行 twitter 用户检索、twitter 内容检索与互联网全网检索。用户界面友好，简单、大方、易使用。很多政企单位都可以使用互联网信息采集系统进行周密的研究工作。

该系统的主要工作如下：

（1）互联网信息采集系统可以对最新推文进行实时抓取与记录；

（2）互联网信息采集系统可以对用户的生活习惯进行周密的分析研究；

（3）互联网信息采集系统通过对 twitter 用户名称进行检索；

（4）互联网信息采集系统通过对 twitter 内容信息进行检索；

（5）互联网信息采集系统通过输入关键字在互联网全网进行检索。

1.4 结构框架

论文分为五章，各章内容组织如下：

第一章介绍了互联网信息采集系统所处的研究环境以及它能为信息收集带来的收益，分别把我国的基本国情与国外情况做对比然后得到解决方法。同时也明确了本文的主要研究方向和要解决的问题。

第二章介绍了互联网信息采集系统的开发平台和运行环境。详细介绍了 twitter 的 Open API 的调用、Smarty 模版引擎、PHP 脚本语言、B/S 和 C/S 模式、三层体系结构，给出了系统软硬件运行环境。

第三章对于互联网信息采集系统功能的可实现方面做了阐述。分别从社会因素和技术因素确定了程序可以正常的运行，且通过分析时下的基本条件确定可以满足程序的各个功能模块的实现。

第四章系统的设计与实现。介绍了系统设计的目的和原则，按照先总体后模块的设计方法对系统进行了设计。针对该系统进行了数据库和服务器选择，并且说明了在程序的具体实现过程中的选择是正确的，程序在安全方面的存在的问题也进行了比较全面的解决与检查。

第五章总体概括与新的思想。从实践中检验了互联网信息采集系统的高效、可靠，而且对应用程序进行了总体概括并开阔了新的思想。

第二章相关技术介绍

2.1 系统调用的 API

2.1.1 OAuth 简介

OAuth 是一个开放式验证协议，是由互联网工程任务组（Internet EngineeringTask Force ，IETF）创建起草。允许用户在不同的网站上共享其私有的数据，而这一过程中网站无需将用户的凭证泄露给第三方。OAuth 增强了网站的安全性，保护了用户的隐私[4]。

2.1.2 OAuth 原理

在 OAuth 的验证过程中，有三个基本元素：用户、服务提供商、使用者。其中其使用者和服务提供商又是主要角色。使用者使用 OAuth 服务提供上的 web 站点和应用，服务提供商允许使用者通过 OAuth 访问其 web 应用程序。

为了抓取 twitter 指定用户的推文及个人信息，首先需要获得一个 twitter 帐户。OAuth 协议认证虽然是用户的登陆过程，但是，其登陆始终是控制在服务提供商所提供的 web 页面内，而并非使用者可以自定义的页面，在为第三方提供方便的同时，保证了用户的隐私不被泄露[5]。

使用 OAuth 1.0 进行认证和授权的过程如下所示:

（1）用户访问使用者的网站，想操作自己存放在服务提供商的资源。

（2）使用者向服务提供商请求一个临时令牌。

（3）服务提供商验证使用者的身份后，授予一个临时令牌。

（4）使用者获得临时令牌后，将用户引导至服务提供商的授权页面请求用户授权。在这个过程中将临时令牌和使用者的回调连接发送给服务提供商。

（5）用户在服务提供商的网页上输入用户名和密码，然后授权该使用者访问所请求的资源。

（6）授权成功后，服务提供商引导用户返回使用者的网页。

（7）使用者根据临时令牌从服务提供商那里获取访问令牌。